

Aging and species abundance in the Tangled Nature model of biological evolution.

Matt Hall,¹ Kim Christensen,² Simone A. di Collobiano,² and Henrik J. Jensen^{1,*}

¹*Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2BZ, U.K.*

²*Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BW, U.K.*

(Dated: 4. Feb. 02)

We present an individual based model of evolutionary ecology. The reproduction rate of individuals characterized by their genome depends on the composition of the population in genotype space. Ecological features such as the taxonomy and the macro-evolutionary mode of the dynamics are emergent properties. The macro-dynamics exhibit intermittent two mode switching with a gradually decreasing extinction rate. The generated ecologies become gradually better adapted as well as more complex in a collective sense. The form of the species abundance curve compares well with observed functional forms. The models error threshold can be understood in terms of the characteristics of the two dynamical modes of the system.

PACS numbers: 05.65.+b, 87.2-n, 87.23.Cc, 87.23.Kg

I. INTRODUCTION

The dynamics and organization of biological ecosystems is a fascinating example of complex interacting systems with many levels of emerging structure and time scales. Biological evolution creates intricate taxonomic hierarchies presumably as an effect of mutation, natural selection and the ensuing adaptation. Taxonomic structures from the level of individuals through species and genera up to kingdoms are generated and vanish again in a never ending succession. Different strata in the hierarchy are described by very different timescales and with very different types of dynamics. At the level of individuals, fairly well defined characteristic lifetimes exist for each specific type (species) and the population dynamics can be considered smooth. This picture changes as one considers the system at the more coarse grained level of species and genera. The lifetime distribution of, e.g., genera is broad (see e.g. [1]) and the dynamics is intermittent [2, 3, 4, 5]. In the spirit of the traditional approach of statistical mechanics it is interesting to consider models, defined at a microscopic level, which are able to reproduce the large scale temporal and taxonomic structures.

In the present paper we consider a model of individuals identified solely by their genome. The model was introduced in Ref. [6] where we also presented a discussion of the qualitative behavior of the model. We combine ecology with evolution by considering interacting individuals which can multiply (sexually or asexually) subject, potentially, to mutations. The size of the total population fluctuates, the average being controlled by the amount of available resources. From these three minimal ingredients emerge segregation in genome space, to be interpreted as the appearance of species, and a complex intermittent dynamics, to be interpreted as extinction and creation

events at the higher taxonomic levels. The entire taxonomic hierarchy is an emergent property of the dynamics at the microscopic level of individuals. We characterize the configurations generated in genotype space in terms of the species abundance curve, and find a good qualitative agreement with the functional form typically found for real ecosystems. The intermittent dynamics is characterized by the statistics of the duration of the quasi-stable epochs or in other words the waiting times between transitions. We find a broad distribution of durations and observe a gradual aging of the macro-dynamics. No stationary state is ever reached.

A. Related Models

Many mathematical models of biological evolution have been developed according to the usual statistical mechanics agenda of generating the macroscopic complex behavior from simplistic microscopic definitions. An elegant review of this endeavor has recently been given by Drossel [7]. Here we limit ourselves to a discussion of similarities and differences between our model and related studies.

Let us first mention models which define the ecosystem in terms of individuals. Higgs and Derrida [8] studied speciation in a model consisting of a fixed number of individuals. Each individual is represented by a genome modeled as a string of zeros or ones, like in Eigen and collaborators' seminal work on quasi-species [9]. Higgs and Derrida demonstrated that a sexually reproducing population breaks up into distinct species when only individuals with a sufficiently similar genome sequence are allowed to produce offspring. This agrees with a large bulk of experimental work [10]. Gavrillets and collaborators [11, 12, 13] have made use of similar models generalized in particular to be able to study geographical and temporal aspects of speciation. These studies differ from ours in assuming a fixed population size and by defining a fitness function for pairs of individuals which is constant if the Hamming distance between the genomes of

*Author to whom correspondence should be addressed:
h.jensen@ic.ac.uk; URL: <http://www.ma.ic.ac.uk/~hjjens/>

two individuals is small enough, and zero otherwise. Our model allows the total size of the population to fluctuate and the fitness of pairs of individuals (or in the asexual case single individuals) depends on the composition of the population at a given instant in time.

It is also important to mention the fitness landscape approach first pioneered by Wright [14, 15], who considered gene frequencies, and was brought to the attention of the statistical mechanics community mainly through Kauffman's so called NK model [16, 17]. The main focus of the NK model, and of the later co-evolutionary NKC model [17], is the study of epistatic interactions (the influence of one gene on another) by use of fitness functions. The main difference between our model and Kauffman's models is that the fitness of an individual in our system depends on the *frequencies* by which other locations in genotype space are occupied.

Taylor and Higgs [18] have studied pleiotropy and epistasis (the influence of one gene on several traits and the influence of one gene on another) in a model that combines and generalizes aspects of the Higgs-Derrida model with the epistatic interactions of Kauffman's models. Taylor and Higgs then derive a phenotypical fitness for the specific genotype. Kaneko and Yomo [19] have also studied models in which the difference between phenotype and genotype is accounted for explicitly. In our model we make the drastic simplification not to distinguish between genotype and phenotype.

Other models consider species as the elementary building block; these models neglect the specifics of the dynamics arising from reproduction and mutations at the level of individuals. The simplest of these models is the Bak-Sneppen model [20]. The model aims to demonstrate that co-evolutionary interactions are sufficient to produce intermittent dynamics which is then related to intermittency in the fossil record and to Eldredge and Gould's concept of punctuated equilibrium [2, 3, 4, 5]. Each species is characterized by a single number between zero and one, the fitness, and the total number of species is kept constant. The model has interesting statistical properties but is difficult to relate to biological evolution.

Species level models of more detail than the Bak-Sneppen model have been formulated recently by McKane, Alonso and Solé [21] and by Drossel, Higgs and McKane [22]. The emphasis in these models is on predator-prey interactions and food-webs and are generalizations of early work by May [23] and May and Anderson [24]. Our model is intended to include all types of interactions between individuals, e.g. antagonistic or collaborative relationships, in addition to predator-prey competitions. Another important difference is that we define our model at the level of individuals in order to be able to study the emergence of species, something not possible in a species based model.

Most models of biological evolution assume that the dynamics is in a statistically stationary state. One marked exception is the model considered by Sibani and collaborators [25, 26, 27, 28]. This is an abstract species

based model consisting of random walks in a rugged fitness landscape. The statistics of the jumps in this landscape are the same as the record statistics considered some time ago by Sibani and Littlewood [29]. The pace of the dynamics of the model gradually slows down as indicated by a logarithmically decreasing extinction rate. As we shall see below our individual based model also exhibits aging, a property found to be consistent with analysis of the fossil record [1].

The paper is organized as follows. In the next section we define the model in detail. In Sec. III we discuss the modes of the models emergent dynamics. In Sec. IV we show how the configurations generated dynamically gradually become better adapted in a collective sense. Sec. V demonstrates that the ecologies generated in the model exhibit characteristics similar to those observed in real ecologies. In Sec. VI we discuss the models ability to address the issue concerning why sexual reproduction can compete evolutionary with asexual reproduction. Sec. VII contains an analysis of the error threshold. We briefly present in Sec. VIII a scan of the behavior of the model for a range of the control parameters and in Sec. IX we conclude and summarize.

II. DEFINITION OF MODEL

We describe here in detail the structure and dynamics of the model which we, with an allusion to Darwin's notion of the *Tangled Bank*, called the Tangled Nature (TaNa) model to stress the model's emphasis on ecological interactions [6].

A. Interaction

We represent an individual by a vector $\mathbf{S}^\alpha = (S_1^\alpha, S_2^\alpha, \dots, S_L^\alpha)$ in genotype space \mathcal{S} . This representation is frequently used, see e.g. Ref. [8, 9, 11, 17, 30]. Here S_i^α may take the values ± 1 , i.e. \mathbf{S}^α denotes one of the corners of the L dimensional hypercube (in the present paper we use $L = 20$). The coordinates S_i^α may be interpreted as genes with two alleles, or a string of either pyrimidines or purines. We think of genotype space \mathcal{S} as containing all possible ways of combining the genomic building blocks into genome sequences. Many sequences may not correspond to viable organisms. Whether this is the case or not is for the evolutionary dynamics to determine. All possible sequences are made available for evolution to select from.

Individuals are labeled by Greek letters $\alpha, \beta, \dots = 1, 2, \dots, N(t)$. When we refer, without reference to a specific individual, to one of the 2^L positions in genome space, we use roman superscripts $\mathbf{S}^a, \mathbf{S}^b, \dots$ with $a, b, \dots = 1, 2, \dots, 2^L$. Many different individuals $\mathbf{S}^\alpha, \mathbf{S}^\beta, \dots$, may reside on the same position, say \mathbf{S}^a , in \mathcal{S} .

The ability of an individual α to reproduce is controlled by $H(\mathbf{S}^\alpha, t)$:

$$H(\mathbf{S}^\alpha, t) = \frac{1}{cN(t)} \sum_{\mathbf{S} \in \mathcal{S}} J(\mathbf{S}^\alpha, \mathbf{S})n(\mathbf{S}, t) - \mu N(t), \quad (1)$$

where c is a control parameter (see below), $N(t)$ is the total number of individuals at time t , the sum is over the 2^L locations \mathbf{S} in \mathcal{S} and $n(\mathbf{S}, t)$ is the occupancy of position \mathbf{S} . Two positions \mathbf{S}^a and \mathbf{S}^b in genome space are coupled with the fixed random strength $J^{ab} = J(\mathbf{S}^a, \mathbf{S}^b)$ which can be either positive or negative or zero. The coupling is non-zero with probability Θ (throughout the paper we use $\Theta = 0.25$), in which case we assume $J^{ab} \neq J^{ba}$ to be a deterministic but erratic function of the two positions \mathbf{S}^a and \mathbf{S}^b . We have checked that the specific details of the form of the distribution of the non-zero values of the function $J(\mathbf{S}^a, \mathbf{S}^b)$ are irrelevant. We choose accordingly a form mainly determined by its numerical efficiency. In the next subsection we describe the details of the specific procedure used. The distribution of the generated interaction strengths is shown in Fig. 3 below.

1. Generation of interaction matrix

The interaction between two locations in genotype space, \mathbf{S}^a and \mathbf{S}^b is generated as a product $J(\mathbf{S}^a, \mathbf{S}^b) = \Theta(\mathbf{S}^a, \mathbf{S}^b)I(\mathbf{S}^a, \mathbf{S}^b)$. The first factor $\Theta(\mathbf{S}^a, \mathbf{S}^b)$ is obtained by interpreting the sequences \mathbf{S}^a and \mathbf{S}^b as binary numbers (letting $-1 \mapsto 0$) and perform the XOR operation on the binary pair to obtain a new integer. This integer is used as an index in a lookup list to obtain either a 0 or 1 as the value of $\Theta(\mathbf{S}^a, \mathbf{S}^b)$. In case 1 is returned, the element of the $I(\mathbf{S}^a, \mathbf{S}^b)$ matrix is obtained in a similar way. This time, however, two arrays are needed. Each auxiliary array is of length 2^L and now the arrays contain uniformly distributed random numbers drawn from the interval $[-1, +1]$. The pair of arrays is necessary in order to reproduce the asymmetry of the $I(\mathbf{S}^a, \mathbf{S}^b)$ matrix. Two indices are generated from the \mathbf{S}^a and \mathbf{S}^b . The first via the same XOR operation used to calculate the $\Theta(\mathbf{S}^a, \mathbf{S}^b)$ matrix element, whereas the second is simply the integer representing \mathbf{S}^b . The strength of interaction is taken to be the product of the members of each array at the appropriate location. This ensures that the elements of the matrices are non-symmetric due to the second array index depending on the order of the operation. This procedure is numerically extremely efficient and deterministic, but has the side effect of generating a distribution of a slightly unusual form, see Fig. 3.

We stress that the coupling matrix $J(\mathbf{S}^a, \mathbf{S}^b)$ is meant to included *all* possible interactions between two individuals of a given genomic constitution. In our simplistic approach, a given genome is imagined to lead uniquely to a certain set of attributes (phenotype) of the individuals/organisms. The locations \mathbf{S}^a and \mathbf{S}^b represent blueprints for organisms that exist *in potentia*. The positions may very likely be unoccupied but, if we were

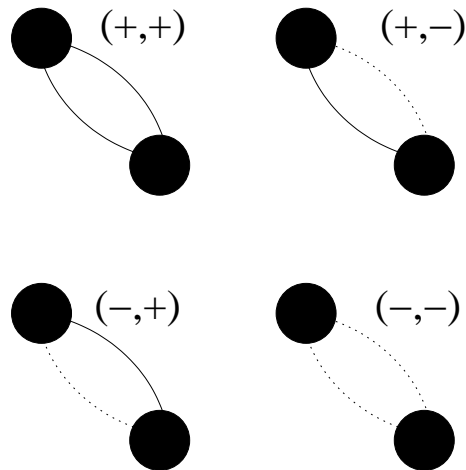


FIG. 1: Examples of possible realizations of the couplings J^{ab} between different positions \mathbf{S}^a and \mathbf{S}^b in genotype space representing collaborative (+,+), antagonistic (-,-) or predator prey (+,-) & (-,+) relationships.

to construct individuals according to the sequences \mathbf{S}^a and \mathbf{S}^b the two individuals would have some specific features. The relationship between an organism of *design* \mathbf{S}^a and one of design \mathbf{S}^b may be as predator and prey or parasitic, i.e. $J^{ab} > 0$ and $J^{ba} < 0$, but it can also be collaborative ($J^{ab} > 0$ and $J^{ba} > 0$) or antagonistic ($J^{ab} < 0$ and $J^{ba} < 0$), see Fig. 1. And certainly in some cases J^{ab} may represent less direct couplings, e.g. some animals may not eat trees, nevertheless they breath the oxygen produced by the rain forest. In order to emphasis co-evolutionary aspects we have *excluded* “self-interaction” among individuals located at the same position \mathbf{S} in genome space, i.e. $J(\mathbf{S}, \mathbf{S}) = 0$ for all $\mathbf{S} \in \mathcal{S}$. It is important to mention that including self-interactions of the same order of strength as the J -couplings does not change the qualitative behavior of the model.

The conditions of the physical environment are simplistically described by the term $\mu N(t)$ in Eq. (1), where μ determines the average sustainable total population size. That is, the total carrying capacity of the environment. An increase in μ corresponds to harsher physical conditions. This is a simplification, though one should remember that what is often considered as the physical conditions, e.g. temperature or oxygen density, is to a degree determined by the activity of other organisms and is therefore really a part of the biotic conditions. Consider, for example, the environment experienced by the bacterial flora in the intestines. Here one type of bacteria live very much in an environment strongly influenced by the presence of other types of bacteria. In this sense some fluctuations in the environment may be thought of as included in the coupling matrix $J(\mathbf{S}^a, \mathbf{S}^b)$.

B. Reproduction, mutations and annihilation

Asexual reproduction consists of one individual being replaced by two copies. Successful reproduction occurs for individuals \mathbf{S}^α with a probability per time unit given by

$$p_{off}(\mathbf{S}^\alpha, t) = \frac{\exp[H(\mathbf{S}^\alpha, t)]}{1 + \exp[H(\mathbf{S}^\alpha, t)]} \in [0, 1]. \quad (2)$$

In the case of sexual reproduction an individual \mathbf{S}^α is picked at random and paired with another randomly chosen individual \mathbf{S}^β with Hamming distance $d = \frac{1}{2} \sum_{i=1}^L |S_i^\alpha - S_i^\beta| \leq d_{max}$ (allowing at most d_{max} pairs of genes to differ). The pair produces an offspring γ with probability $\sqrt{p_{off}(\mathbf{S}^\alpha, t)p_{off}(\mathbf{S}^\beta, t)}$, where S_i^γ is chosen at random from one of the two parent genes, either S_i^α or S_i^β . For $d_{max} \geq 1$ this procedure may be thought of as being similar to recombination. The maximum separation criterion has been studied by several authors, see e.g. [8, 11].

We allow for mutations in the following way: with probability p_{mut} per gene we perform a change of sign $S_i^\gamma \rightarrow -S_i^\gamma$, during the reproduction process.

For simplicity, an individual is removed from the system with a constant probability p_{kill} per time step (we use $p_{kill} = 0.2$). This procedure is implemented both for asexual and sexually reproducing individuals.

A time step consists of *one* annihilation attempt followed by *one* reproduction attempt. One generation consists of $N(t)/p_{kill}$ time steps, which is the average time taken to kill all currently living individuals.

Initially we place $N(0) = 500$ individuals at randomly chosen positions. The results are independent of initial conditions. We obtain the same results if all individuals are located at the same position initially.

The present paper's main focus is on the asexual mode of reproduction and results presented are for asexual individuals except otherwise stated. For completeness and for comparison we do, however, in Secs. VI and VIII consider sexual reproduction, simplicity which is defined as follows.

III. DYNAMICAL STABILITY

Neglecting fluctuations in the occupancy $n(\mathbf{S}, t)$ the above dynamics is described by the following set of equations (one equation for each position in the genotype space):

$$\begin{aligned} n(\mathbf{S}, t+1) = & n(\mathbf{S}, t) \\ & + \{p_{off}(\mathbf{S}, t)[2(1 - p_{mut})^L - 1] - p_{kill}\} \frac{n(\mathbf{S}, t)}{N(t)} \\ & + 2p_{mut}(1 - p_{mut})^{L-1} \sum_{\langle \mathbf{S}', \mathbf{S} \rangle} p_{off}(\mathbf{S}', t) \frac{n(\mathbf{S}', t)}{N(t)}, \quad (3) \end{aligned}$$

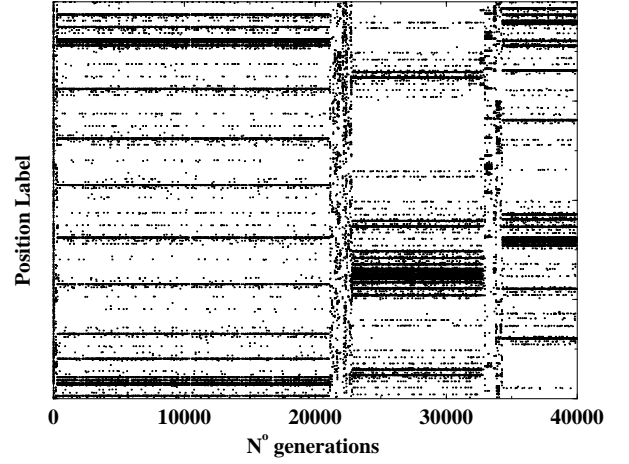


FIG. 2: The occupation in genotype space plotted as a function of generation time. The genotypes are enumerated in an arbitrary manner. If a position is occupied at a given moment in time a dot is placed at the corresponding number along the y-axis at that instant in time. Parameters are $c = 0.5$, $\mu = 0.005$ and $Lp_{mut} = 0.25$.

where the sum is over the nearest neighbors of \mathbf{S} . Stationary solutions require the system to find configurations in genotype space for which all positions satisfy the demand that either $n(\mathbf{S}, t) = 0$ or if $n(\mathbf{S}, t) \neq 0$ (neglecting the mutational back flow represented by the last term in Eq. (3)), we must have

$$p_{off} = \frac{p_{kill}}{2(1 - p_{mut})^L - 1} \equiv p_{q-ESS} \quad (4)$$

The fitness $p_{off}(\mathbf{S}^a, t)$ of individuals at a position \mathbf{S}^a depends on the occupancy $n(\mathbf{S}^b, t)$ of all the sites \mathbf{S}^b with which site \mathbf{S}^a is connected through couplings J^{ab} . Accordingly, a small perturbation in the occupancy at one position may be able to disturb the balance in Eq. (4) between $p_{off}(\mathbf{S}, t)$, p_{kill} and p_{mut} on connected sites. In this way an imbalance at one site can spread as a chain reaction through the system, possibly causing a global reconfiguration of the occupancy in genotype space.

We show in Fig. 2 the occupancy in genotype space plotted as a function of time for asexual reproduction. Periods of stable configurations are separated by fast transitions. We have called the stable periods “quasi-Evolutionary Stable Strategies” or q-ESS since they are reminiscent of the Evolutionary Stable Strategies (ESS) introduced by Maynard Smith [31].

It is interesting to investigate just how stable the q-ESS are. We have done this by applying different types of perturbations in the q-ESS. The result is that the q-ESS are very stable against global perturbations such as a brief or a lasting increase in control parameters μ , c or p_{kill} . Changes of up to 50% in these parameters, either permanently or for a period of 100 generations, only effect the total population size and is typically not able to kick the population out of its present q-ESS configuration in

genotype space. In contrast, a similar perturbation of the mutation rate easily destabilizes the q-ESS configuration.

We stress that the segregation (or speciation) to be discussed below is an effect of different couplings between different positions \mathbf{S}^a and \mathbf{S}^b . When we assume $J(\mathbf{S}^a, \mathbf{S}^b) = J_0$ independent of \mathbf{S}^a and \mathbf{S}^b , the population is not concentrated around a subset of the positions in genotype space, instead the population is smeared out through the space in a diffuse manner. Self-interaction, however, can cause segregation in a rather trivial way. Namely, if we include a distribution of $J(\mathbf{S}, \mathbf{S})$ values, segregation may occur even in the case where all interaction terms assume the same value: $J(\mathbf{S}^a, \mathbf{S}^b) = J_0$ for $\mathbf{S}^a \neq \mathbf{S}^b$. However, this type of selection of configurations in genotype space is not very interesting since the sites to become occupied is determined by the arbitrarily assigned self-interactions $J(\mathbf{S}, \mathbf{S})$ and not by the collective dynamical adaptation at play when $J(\mathbf{S}, \mathbf{S}) = 0$ and $J(\mathbf{S}^a, \mathbf{S}^b)$ assumes a distribution of different J -values. In reality one will expect the selection of species to be caused by a mixture of self-interaction and interaction between different species. To decide which one is dominant might be difficult and will certainly be system specific.

There is a significant difference between the distribution of active couplings, $p_{act}(J(\mathbf{S}^a, \mathbf{S}^b))$, in the q-ESS and the distribution during the hectic transitions. We show in Fig. 3 the distribution from which the $J_{bare}(\mathbf{S}^a, \mathbf{S}^b)$ are sampled together with the distribution of couplings between occupied sites after a large number of generations. During the hectic phases there is clearly no noticeable difference between the “bare” distribution of the $J(\mathbf{S}^a, \mathbf{S}^b)$ and the distribution of active couplings, i.e. couplings between occupied positions. During the q-ESS we observe a slight bias towards positive J -values of the active couplings. This slight shift towards more positive couplings will, according to Eqs. 1 and 2, lead to an increased reproduction rate during the q-ESS. The manifestation of this difference between the hectic periods and the q-ESS is illustrated in Fig. 4. We see that the distribution of H -values in the q-ESS during both modes of the dynamics contains a narrow peak. In the q-ESS, the peak in $p(H)$ is separated from a strongly negative band of support. The values of H in this band are so negative that the corresponding $p_{off}(H)$ are negligible (see the insert in Fig. 4). Genotype positions corresponding to this band consist of unfit positions next to highly occupied and very fit positions. The reason these positions are occupied at all is that they are supplied by mutations occurring on the neighboring fit positions. The conclusion of these considerations is that the dynamics during the q-ESS as well as during the hectic periods are controlled by the reproduction of individuals with H -values in the two respective peaks of $p(H)$.

The location of the peaks of $p(H)$ is determined in the following way. During the hectic periods the occupation of positions in genotype space is highly unstable and $n(\mathbf{S}, t + 1)$ is only related to $n(\mathbf{S}, t)$ in an erratic

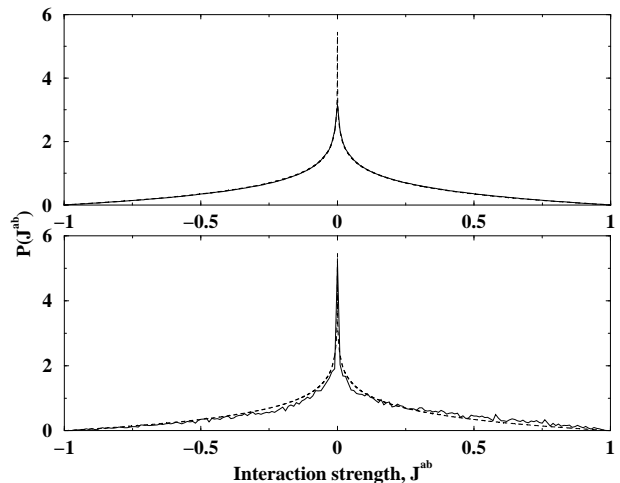


FIG. 3: The distribution from which the values of the couplings $J(\mathbf{S}^a, \mathbf{S}^b)$ are drawn at the start of the simulations (dashed curve) together with the probability density function of the couplings between occupied sites (solid curve) during the hectic periods (top panel) and during the q-ESS (bottom panel). Parameters are $c = 0.01$, $\mu = 0.01$ and $Lp_{mut} = 0.2$.

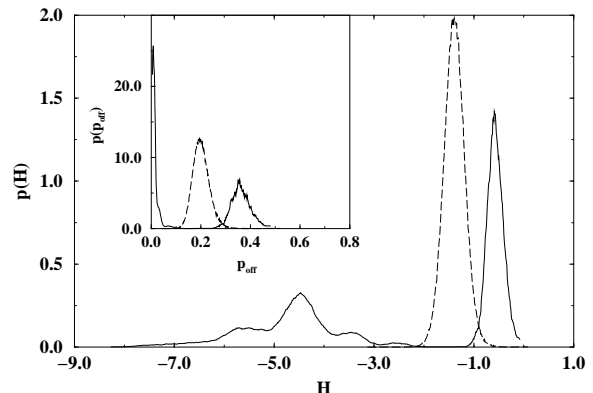


FIG. 4: The probability density function for the weight function H (main frame) and reproduction rates p_{off} (insert) during the hectic transitions (dashed curve) and in the q-ESS (solid curve). Parameters are $c = 0.08$, $\mu = 0.005$ and $Lp_{mut} = 0.25$.

way, the balance equation Eq. (3) is never fulfilled for nonzero $n(\mathbf{S}, t) = 0$. The only constraint on p_{off} during the hectic periods is accordingly that the total population remains constant on average which implies that on average $p_{off} = p_{kill}$. This explains why in Fig. 4 the peak in $p(H)$ during the hectic periods corresponds to a peak in $p(p_{off})$ centered at $p_{kill} = 0.2$. The situation is different during the q-ESS. Here the occupation of the selected positions in genotype space remains approximately constant and Eq. (4) applies. Substituting the relevant values $p_{kill} = 0.2$, $p_{mut} = 0.0125$ and $L = 20$ into Eq. (4) produces $p_{off} = 0.36$ which explains the position of the peak in $p(p_{off})$ during the q-ESS.

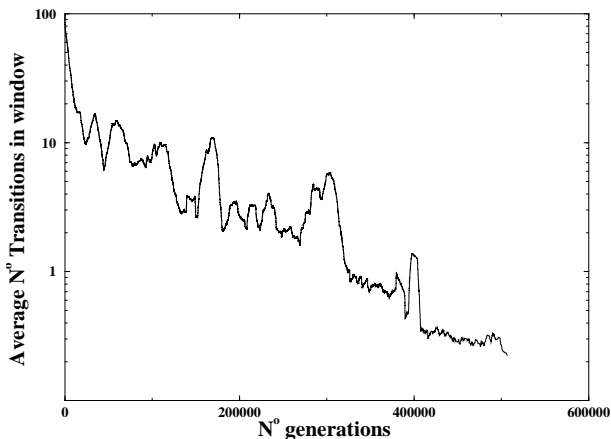


FIG. 5: The average number of transitions during a window of size $T = 1000$ generations as a function of generation time. Parameters are $c = 0.01$, $\mu = 0.01$ and $Lp_{mut} = 0.2$. The average is over 400 realizations.

IV. AGING

For simplicity we concentrate again on the asexual model in this section. For large genome length L the system is always in a transient. The time needed to reach the stationary state increases exponentially with L and is therefore unreachable for any biologically relevant values of L .

A. Increasing q-ESS durations

The gradual change in the statistical measures of the model is seen directly as a slow increase with time of the average duration of the q-ESS. To demonstrate this we show in Fig. 5 the average number of transitions $\Omega_T(t)$ between q-ESS within a time window of fixed size T as a function of time t measured in number of generations. It is clear that $\Omega_T(t)$ decreases with increasing t , however it is very difficult to obtain sufficient statistics to be able to determine the functional dependence of $\Omega_T(t)$ on t , though a very slow exponential t dependence is suggested by Fig. 5. Despite these sampling difficulties, it is evident that the duration of the q-ESS, on average, increases with time. This corresponds to a decrease in the extinction rate, consistent with analysis of the fossil record [1].

B. Increasing population size, diversity and complexity

The gradual growth of the duration of the stable q-ESS epochs indicates that the dynamics of the system is able to produce more stable or better adapted configurations in genotype space. It is difficult to test quantitatively the stability of the q-ESS with respect to perturbations. That the population is distributed in an in-

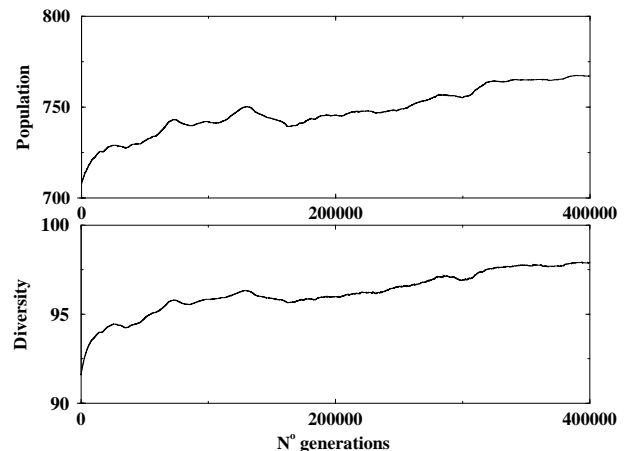


FIG. 6: The ensemble averaged total population (top) and diversity (bottom) as function of generation time for the same ensemble as in Fig. 5.

creasingly more efficient manner in genotype space can be seen directly from the increase in the total population size $N(t)$ averaged over an ensemble of different realizations of the stochastic elements of the dynamics. Fig. 6 contains the average total population $\langle N(t) \rangle$ together with the ensemble average of the diversity $\langle D(t) \rangle$, where $D(t)$ is defined as the number of different occupied positions \mathbf{S} in genotype space at time t . The average diversity also increases with time.

Let us briefly consider how the total population size can increase. We saw in Sec. III that essentially p_{off} is narrowly distributed either about p_{kill} or, as in the q-ESS, about the value p_{q-ESS} in Eq. (4). The increase in $N(t)$ is therefore not an effect of a gradual increase in p_{off} . Simulations indeed confirm that the average offspring probability always remains constant over the entire run. In biological observations and experiments, for reasons of uniqueness, the reproduction rate is identified as *fitness*. In this sense the fitness of the individuals remains, on average, constant in the TaNa model, as presumably is also the case in biological macroevolution; though the microbial experiments by Lenski [32] demonstrate that the reproductive fitness can increase as a result of adaptation in microevolution.

The increase in the average population size $\langle N(t) \rangle$ observed in the TaNa model is caused by the system ability to generate configurations that increase the interaction term in the weight function $H(\mathbf{S}, t)$ defined in Eq. (1). When the first term increases, the second term $\mu N(t)$ in Eq. (1) can increase as well, while the total $H(\mathbf{S}, t)$ remains on, average, fixed.

The increase in the interaction term of $H(\mathbf{S}, t)$ is achieved in several ways. Firstly, the population is spread out onto an increasing number $D(t)$ of different genotypes, as seen in Fig. 6. Moreover, the evolutionary dynamics tends to produce occupied sites which are interacting with an increased number of other occupied sites, i.e., the number of non-zero terms $J(\mathbf{S}^\alpha, \mathbf{S})n(\mathbf{S})$ in

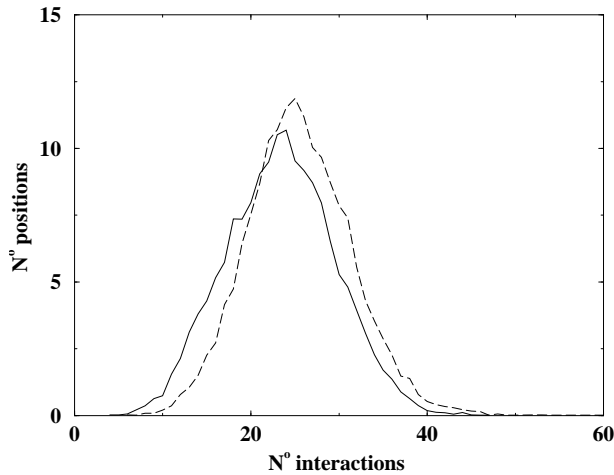


FIG. 7: The number of occupied positions in genotype space with a given number of active links connected to other occupied position in genotype space. The solid line is after 500 generations, the dashed line is recorded after 10^5 generations. Parameters are $c = 0.01$, $\mu = 0.01$ and $Lp_{mut} = 0.2$.

$H(\mathbf{S}^a, t)$ in Eq. (1) grows as the system produces configurations that are able to benefit better from the possible mutual interactions represented by $J(\mathbf{S}^a, \mathbf{S})$. The distribution of active interaction links is shown at an early and a much later time in Fig. 7. We have not been able to resolve a shift with time in the distribution of the values of the active interaction strengths $J(\mathbf{S}^a, \mathbf{S}^b)$.

The increase in the diversity and the number of active links connected to an occupied position can be interpreted as an increase in the complexity of the configurations produced by the evolutionary dynamics. Selection and adaptation operate at the level of the entire configuration in genotype space rather than at the level of individual genotypes. This highlights that the biological concept of fitness makes most sense when considered as a *collective property of an ecology*, rather than an observable characteristic of the individual species or individual members of a population.

C. Record Statistics

The observation in the previous section that the dynamics of the TaNa model leads to an increase in a number of measurable quantities, taken together with the intermittent nature of the dynamics, suggests that the transitions between consecutive q-ESS epochs correspond to record transitions. One can imagine that some characteristic measures of the collective level of adaptation of the configurations generated in genotype space achieves an ever increasing value as the system undergoes a transition from one q-ESS to the next.

Sibani and collaborators [25, 26, 27, 28, 29] have studied record dynamics and shown that the probability for n records in a sequence of t independently drawn ran-

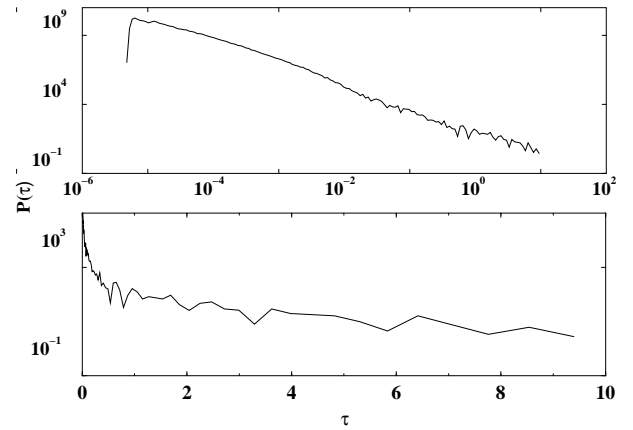


FIG. 8: The probability density $p(\tau)$ of the logarithmic waiting time τ . The top graph is a double logarithmic plot of $p(\tau)$ exhibiting a power law behavior in the region of small τ values. The bottom plot is a linear-log plot of the same data. Here one sees that the behavior of $p(\tau)$ for large values of τ is consistent with a slow exponential decay. Parameters are as for Fig. 7.

dom numbers is Poisson distributed on a logarithmic time scale, or equivalently: that the logarithm of the ratio of the time between the k -th and the $(k-1)$ -th record, $\tau_k = \ln(t_k/t_{k-1})$, is exponentially distributed: $P(\tau > x) = \exp(-\lambda x)$. Sibani and coworkers [25, 26, 27, 28] have also demonstrated the relevance of record statistics to the dynamics of the Kauffman NK model [16, 17].

Accordingly, it is interesting to investigate if the aging observed in the TaNa model exhibits signs of record statistics. To do this, we study the distribution of the variable $\tau_k = \ln(t_k/t_{k-1})$ where t_k denotes the time at which the k -th transition between consecutive q-ESS epochs occurs. We show in Fig. 8 that τ_k is exponentially distributed for large values and algebraically distributed for small values of τ_k .

The exponential tail in Fig. 8 suggests that the transition times in the TaNa model follow record statistics in the region of large τ values, corresponding to the regime of long q-ESS durations. The algebraic form of $p(\tau)$ for small τ values indicates significant correlations for transitions that occurred in rapid succession. The question is which quantity evolves according to record dynamics. We have so far not been able to identify a variable of the system which jumps monotonously to ever higher values at the transition times. In the evolution models studied by Sibani *et al.* [25, 26, 27, 28] the fitness increases through consecutive records. As mentioned above in the TaNa model the reproductive fitness p_{off} remains, on average, constant. The increase in the average duration of the q-ESS (see Fig. 5) suggests that the stability of the configurations in genotype space gradually increases. To explore this, one should study the temporal behavior of the eigenvalue spectrum of the stability matrix of the effective evolution equations in (3). We expect that the number of unstable directions, on average, decreases with

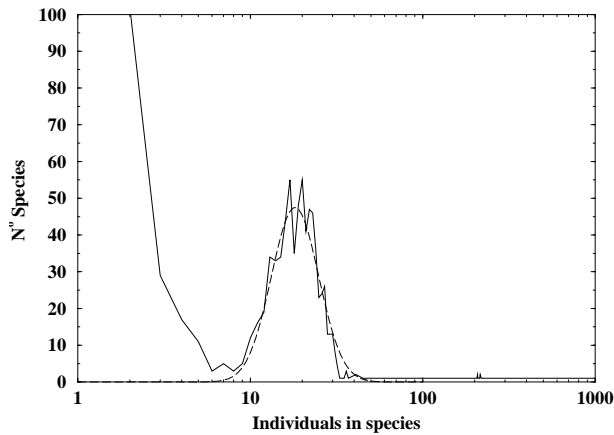


FIG. 9: The species abundance distribution. The peak in the distribution is compared with the lognormal form (dashed curve). Parameters are $c = 0.01$, $\mu = 5 \cdot 10^{-5}$ and $Lp_{mut} = 0.2$.

time, though for a given realization fluctuations probably prevent a strictly monotonous behavior.

V. SPECIES ABUNDANCE FUNCTION

The fundamental quantity to describe an ecology is the species abundance function [33]. The species abundance $W(\rho)$ is the ratio W of species which contains a ratio ρ of the total population. We plot in Fig. 9 the species abundance function for the TaNa model during a q-ESS, in this case we use the term species to denote individual positions in genotype space. A large number of positions are occupied by a small number of individuals; the occupancy of these positions is never established for extended periods during the q-ESS. The robust species contain a reasonable number of individuals and are distributed according to the broad peak. The peak can be fitted by a log-normal curve in a way similar to observed species abundance functions, see e.g. [33]. We note that comparable species abundances functions are found in the predator-prey model studied by McKane, Alonso and Solé [21].

VI. COMPETITION BETWEEN SEXUAL AND ASEXUAL REPRODUCTION

We now consider the competition within the TaNa model between sexual and asexual reproduction in simulations of mixed populations. This is done by adding an extra *reproduction determining* gene to the genome (making $L = 21$ in this section). This extra gene does not explicitly enter in $H(\mathbf{S}, t)$ but dictates an individual's reproductive mode. Mutations to this gene occur during reproduction in the normal way. Obviously, the simplistic nature of the TaNa model excludes the model

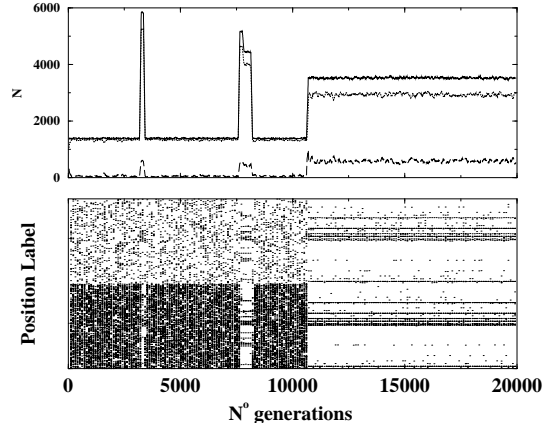


FIG. 10: Run with mixed population. Top frame contains the size of the total population (solid curve), the asexual population (dotted curve) and the sexually reproducing subpopulation (dashed curve). The frame below shows the occupation of the genotypes for the sexually reproducing subpopulation (top half) and the asexual subpopulation (bottom half). Parameters are $c = 0.08$, $\mu = 0.001$, $Lp_{mut} = 0.3$ and $d_{max} = 1$.

from realistically representing all biological features of the difference between sexual and asexual reproduction. Perhaps the most essential difference between sexual and asexual reproduction is the reshuffling of genes caused by the crossing over and recombination involved in sexual reproduction [34]. It is possible that this effect can be captured by our definition of sexual reproduction for $d_{max} \geq 1$. To make the two reproductive modes as similar as possible in all aspects, except for the mixing of parent genes in the sexual case, we redefine in this section slightly the reproduction procedures in the following way. The only difference compared with the definitions in Sec. II B is that an asexual reproducing individual produces *one* new individual and we leave the parent in the system. In sexual reproduction the reproduction probability is assumed to be p_{off} of the first of the two selected parents.

According to Weismann, sexual reproduction is more efficient to adapt to changed conditions because recombination produces a larger variation of types upon which Natural Selection is able to act, see Burt's excellent review [34]. It follows that in a harsher environment, sexual reproduction will be superior to asexual; a phenomenon which is observed [34]. The TaNa model in its present form is not entirely able to account for these facts. It is informative, nevertheless, to investigate the behavior of the model and to understand why it fails in this particular respect.

In Fig. 10 we show the total size of the population together with the sizes of the sexual and asexually reproducing subpopulations. We see that the asexual population is always largest and even more so during the q-ESS.

In Fig. 11 we show the ratio of the average size N_a of

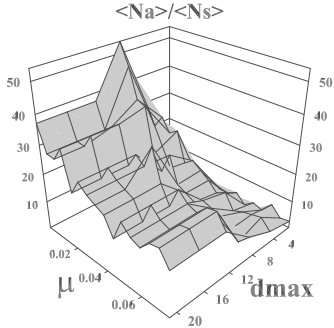


FIG. 11: The ratio of the average number of asexual and sexually reproducing individuals as a function of d_{max} and μ . Parameters are as in Fig. 10.

the asexual reproducing population and the average size N_s of the sexually reproducing population. One notices that the asexual population is always more numerous, but less so for large d_{max} and low μ . That is the region where recombination is most important and the environment (modeled by μ) least friendly. So in this respect the TaNa model exhibit a trend in the right direction in the sense of Weismann, see Burt's review [34].

The reason the sexually reproducing population is unable to out-compete the asexual in the TaNa model is easy to understand. The offspring produced by two sexually reproducing individuals is likely to end up at a position in genotype space different from the parent positions. This effect makes it difficult to maintain the occupancy of the parent positions, even if these positions are highly fit. It is obviously this variation that is expected to make sexual reproduction more efficient in searching genotype space. However, in its present form, the TaNa model allows very large variations in the weight function H in Eq. (1), even for positions very close by in genotype space. Whether this is reasonable or not depends on the interpretation of the genomic sequence \mathbf{S} . If we think of each "gene" S_i as representing, in an averaged way, a section of length $1/L$ of the entire genome, then a separation by Hamming distance one is very significant for L as small as 20; and it is reasonable that two positions in genotype space of this or larger separation may have very different weight functions H . On the other hand, this interpretation means that two individuals being different by a single Hamming distance differ by a major fraction of the entire genome and should therefore not belong to the same species and accordingly not be able to produce any (viable) offspring. We conclude that a more realistic study of the competition between sexual and asexual reproduction by use of the TaNa model should be possible if a more slowly varying weight function is used. Future studies will focus on this problem.

VII. THE ERROR THRESHOLD

At sufficiently large mutation rates p_{mut} , offspring are so different from their parents that the occupation in genotype space rapidly moves from one position to the next. When this happens, it becomes impossible to establish the q-ESS seen in Fig. 2 and consequently the entire simulation consists of one hectic period. The change from the behavior depicted in Fig. 2, where the hectic periods are of much shorter duration than the q-ESS, to the behavior where the q-ESS are absent, occurs over a very narrow region of p_{mut} values. Considering first large values of p_{mut} we gradually decrease p_{mut} in the simulations and we identify the threshold value, p_{th} of p_{mut} at which q-ESS are observed as the error threshold [7, 9]. In Fig. 12 we plot the simulated value of p_{th} for different values of the width parameter c .

We can estimate the c dependence of p_{th} by the following argument. From Fig. 4 we know that the distribution of p_{off} in the hectic periods is centered about p_{kill} and in the q-ESS is centered about p_{q-ESS} defined in Eq. (4). Changing the parameter c will change the width of the distribution of the p_{off} values (see Eqs. (1) and Eq. (2)). It will be possible to establish q-ESS in between the hectic periods if $p_{kill} + \sigma_p \leq p_{q-ESS}$, where σ_p is the half width of the peak in the distribution of p_{off} in the hectic periods. We translate this argument to the distribution of the H values and obtain the following estimate for p_{th} :

$$p_{th} = 1 - 2^{-1/L} [(1 - p_{kill})e^{-\alpha/c} + 1 + p_{kill}]^{1/L}. \quad (5)$$

Here we have assumed that the width σ_p of the peak in the distribution of H values will be given by $\sigma_p = \alpha/c$ (see Eq.(1)) in which case α is a measure of the standard deviation of the factor in Eq. (1) multiplying $1/c$. We have used $\alpha = 0.07$ to fit the simulation data in Fig. 12. This value is somewhat larger but of the right order of magnitude as the corresponding quantity measured during the simulation.

VIII. PARAMETER DEPENDENCE

For completeness, we present here the dependence on the parameters c and μ in the Hamiltonian given in Eq. (1).

We show in Figs. 13 and 14 the averaged occupation measured as the ratio between the average number of individuals and the average number of occupied positions in genotype space for purely asexual and sexual reproducing populations respectively. As expected, the system is able to support the largest populations in the region of small μ parameter and broad distribution of coupling strength, i.e., small values of c . The sexual reproduction is most sensitive to a decrease in the carrying capacity (increase in μ) or a decrease in the width of the range of possible $J(\mathbf{S}^a, \mathbf{S}^b)$ couplings (increase in c).

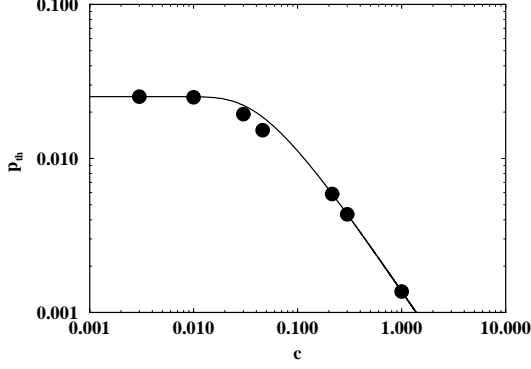


FIG. 12: The loss of q-ESS occurs for mutation rates above the circles. For comparison, the theoretically predicted error threshold $p_{mut}^{th}(c)$ is shown for $\alpha = 0.07$ (see main text). The carrying capacity parameter is $\mu = 0.005$.

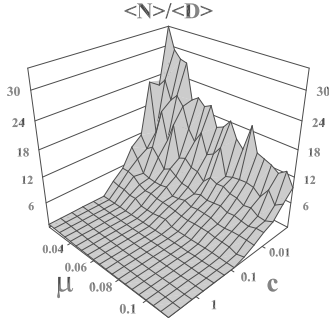


FIG. 13: The ratio between the average size of the population and the average diversity as function of the width parameter c and the physical environment parameter μ . The data are for a system with asexual reproduction with $Lp_{mut} = 0.2$.

IX. DISCUSSION AND CONCLUSION

The Tangled Nature model may be considered as a mathematical framework for the study of evolutionary ecology. The dynamics of the model is defined at the level of individuals, either as asexual or as sexually reproducing individuals. All ecological structure in the model arises through emergence. The model is able to generate many of the observed features of biological evolution starting from a basic implementation of the key assumptions in the Darwinian evolution paradigm.

The density of individuals in genotype space segregates corresponding to the emergence of distinct species. The interaction between individuals gives rise to a jerky or intermittent macro-dynamics, in which quasi-stable configurations (the q-ESS) in genotype space are abruptly

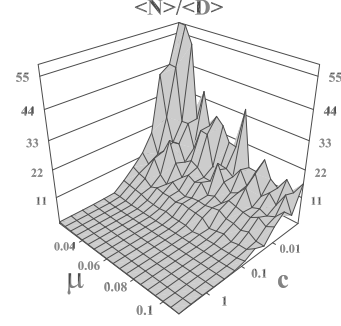


FIG. 14: The same data and parameters as in Fig. 13 except that the data in this figure is for a sexually reproducing population.

replaced by new quasi-stable configurations. This mode of operation can be compared with the intermittent behavior observed in the fossil record and emphasized by Gould and Eldredge's term "punctuated equilibrium" [4]. The TaNa model is always in a transient where the configurations generated as a result of adaptation to the co-evolutionary selective pressure gradually produce configurations, or ecologies, in genotype space which collectively exhibit a higher degree of adaptation, in the sense that the average lifetime of these q-ESS increases slowly. This behavior compares well with the observation that the fossil record indicates a decrease in the extinction rate [1]. The increase in the lifetime of the q-ESS is associated with an increase in the complexity (species diversity, number of active interactions) of successive configurations. This gradual aging together with the intermittent nature of the dynamics suggest that some characteristic of the evolving ecosystem might be undergoing record statistics in the sense of Sibani and co-workers [25, 26, 27, 28]. So far we have unfortunately not been able to identify the appropriate effective variable which moves through the records but we expect this variable to be related to the stability matrix of the effective dynamical equation.

The species abundance distribution generated by the TaNa model encourage future studies of larger populations with longer genome sequences. This will enable a hierarchical study of the taxonomic organisation of the generated ecologies. Using distance criteria in genotype space one can study the clustering of individuals into species, of species into genera, etc. Future studies will also examine the phylogenetic structures in detail, especially during the radiation of species encountered in the transition periods between q-ESS. Using longer genome sequences and a more smoothly varying weight function, we expect the TaNa model to be able to illuminate the evolutionary competition between sexual and asexual reproduction.

Acknowledgments

It is a pleasure to acknowledge very helpful discussions with Paolo Sibani. MH and SAC are supported

by EPSRC studentships and KC gratefully acknowledges the financial support of EPSRC through Grants No. GR/R44683/01 and GR/L95267/01. We thank Paul Anderson for reading the manuscript.

-
- [1] M. E. J. Newman and P. Sibani, Proc. R. Soc. Lond. B **266**, 1593 (1999).
 - [2] N. Eldredge, Evolution **25**, 156 (1971).
 - [3] S. J. Gould, Palaeobiology **3**, 135 (1977).
 - [4] S. Gould and N. Eldredge, Palaeobiology **3**, 114 (1977).
 - [5] N. Eldredge and S. Gould, Nature **332**, 211 (1988).
 - [6] K. Christensen, M. Hall, A. di Collobiano, and H. J. Jensen (2002), arXiv:cond-mat/0104116, To appear in J. Theor. Biol.
 - [7] B. Drossel, Adv. Phys. **50**, 209 (2001).
 - [8] P. G. Higgs and B. Derrida, J. Mol. Evolution **35**, 454 (1992).
 - [9] M. Eigen, J. McCaskill, and P. Schuster, J. Phys. Chem. **92**, 6881 (1988).
 - [10] E. E. Hoster, Evolution **47** (1993).
 - [11] S. Gravilets, Am. Nat. **154**, 1 (1999).
 - [12] S. Gavrillets, H. Li, and M. D. Vose, Proc. R. Soc. Lond. B **265**, 1483 (1998).
 - [13] S. Gavrillets, H. Li, and M. D. Vose, Evolution **54**, 1126 (2000).
 - [14] S. Wright, Proc. 6. Int. Congress of Genetics **1**, 356 (1932).
 - [15] S. Wright, Am. Nat. **131**, 115 (1988).
 - [16] S. A. Kauffman and S. Levine, J. Theor. Biol. **128**, 11 (1987).
 - [17] S. Kauffman, *The Origins of Order* (Oxford University Press, Oxford, 1993).
 - [18] C. F. Taylor and P. G. Higgs, J. Theor. Biol. **203**, 419 (2000).
 - [19] K. Kaneko and T. Yomo, Proc. R. Soc. Lond. B **267**, 2367 (2000).
 - [20] P. Bak and K. Sneppen, Phys. Rev. Let. **71**, 4083 (1993).
 - [21] A. McKane, D. Alonso, and R. V. Solé, Phys. Rev. Let. **62**, 8466 (2000).
 - [22] B. Drossel, P. G. Higgs, and A. J. McKane, J. Theor. Biol. **208**, 91 (2001).
 - [23] R. M. May, Nature **238**, 413 (1972).
 - [24] R. M. May and R. M. Anderson, Proc. R. Soc. Lond. B **219**, 281 (1983).
 - [25] P. Sibani, M. R. Schmidt, and P. A. m, Phys. Rev. Let. **75**, 2055 (1995).
 - [26] P. Sibani, Phys. Rev. Let. **79**, 1413 (1997).
 - [27] P. Sibani and M. Brandt, Int. J. Mod. Phys. B **12**, 361 (1998).
 - [28] P. Sibani and A. Pedersen, Europhys. Let. **48**, 346 (1999).
 - [29] P. Sibani and P. Littlewood, Phys. Rev. Let. **71**, 1482 (1993).
 - [30] H. Wagner, E. Baake, and T. Gerische, J. Stat. Phys. **92**, 1017 (1998).
 - [31] J. Maynard Smith, *Evolution and the theory of games* (Cambridge University Press, 1982).
 - [32] R. E. Lenski and M. Travisano, Proc. Natl. Acad. Sci. USA p. 6808 (1994).
 - [33] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton University Press, Princeton, 2001).
 - [34] A. Burt, Evolution **54**, 337 (2000).